ECONOMETRIC DETERMINANTS OF RESIDENTIAL SOLAR DEVELOPMENT IN THE UNITED STATES

by

Micah D. Thomas

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Economics

Charlotte

2020

Approved by:

_____
Dr. Peter Schwarz

_____
Dr. Craig A. Depken II

_____
Dr. Louis H. Amato

ABSTRACT

MICAH D. THOMAS. Econometric determinants of residential solar development in the United States. (Under the direction of DR. PETER SCHWARZ)

I perform a comprehensive econometric analysis of the demographic determinants of residential solar installations in the contiguous United States. The exponential growth of residential solar in the United States has been aided by declining component costs, environmental concern, and government incentives. Contentious debate surrounds solar incentives as there is uncertainty over their efficacy and equity. Previous research has attempted to provide context to this debate by modeling the development of residential solar in the United States, but it has relied upon partial datasets and Generalized Linear Models. I utilize the comprehensive DeepSolar database on residential solar in the US and, contrary to previous research, find that a Cragg Hurdle model specification outperforms the Generalized Linear Model frameworks. Results of this model suggest that solar incentives have had a positive impact on residential solar development and that residential solar adoption in the United States has been pursued by counties that are characterized by the middle class. This suggests that solar incentives have been both more effective and perhaps more equitable than previously thought.

ACKNOWLEDGMENTS

Thank you to my patient wife Mikala. This thesis stands as a testament to the time that I owed to you, but spent in academic studies. Thank you for your kind encouragement and stoicism throughout this endeavor. Your warm love and cheerful humor sustained my sanity.

Thank you to my parents, Phillip and Cynthia. Your love and encouragement regardless of my self-deprecation has kept me motivated.

I extend my sincerest thanks to my thesis chair, Dr. Schwarz. I greatly appreciate your patience and guidance over the past several years. Your mentorship has been invaluable in developing my understanding of economics.

Thank you to Dr. Amato for your continuing support. You have been a great teacher and mentor and I will always be grateful for what you have done for me.

Thank you to Dr. Depken for all of your time. Your classroom instruction and academic demeanor has left a positive impact on my growth as a student. Your approach to economics was both refreshing and inspiring and I will not shortly forget it.

TABLE OF CONTENTS

LIST OF TABLES vi
LIST OF FIGURES vii
CHAPTER 1: INTRODUCTION 1
CHAPTER 2: DESCRIPTION AND DATA 7
CHAPTER 3: MODEL SELECTION 14
3.1 Dependent Variable Characteristics 14
3.2 Generalized Linear Models (GLM) 17
3.2.1 Generalized Linear Model Framework 17
3.2.2 Poisson Regression Model (PRM) 18
3.2.3 Negative Binomial Regression Model (NBRM) 19
3.2.4 Zero-Inflated GLM Framework 20
3.3 Censored Regression Models (CRM) 21
3.3.1 Tobit Regression Model 21
3.3.2 Cragg Hurdle Model 21
3.3.3 Heckman Two-step Regression Model 23
3.3.4 Stochastic Frontier Analysis 23
3.4 Overdispersion Testing 23
3.5 Model Performance 26
CHAPTER 4: RESULTS 27
4.1 First Stage Selection Model 29
4.2 Second Stage Model 29
4.3 Complete Model 30
CHAPTER 5: CONCLUSIONS 32
REFERENCES 34

LIST OF TABLES

TABLE 1: Variable Descriptions and Sources 8
TABLE 2: Summary Statistics 9
TABLE 3: Overdispersion Tests for the PRM 24
TABLE 4: Dispersion Values for the GLM Methods 25
TABLE 5: Performance Statistics of GLMs and CRMs 26
TABLE 6: Cragg Hurdle Model Results 28
TABLE 7: Average Marginal Effects of the Cragg Hurdle Model. 30

LIST OF FIGURES

FIGURE 1. Histogram of the number of residential solar installations in the US 15
FIGURE 2. Residuals vs fitted values of the OLS regression.....16
FIGURE 3. Dispersion Comparison of the GLM Methods.....25

CHAPTER 1: INTRODUCTION

Over the past several decades, residential solar saw tremendous growth in its development and adoption in the United States. Attributed largely to economies of scale and advances in materials and manufacturing, the cost of a residential system has decreased dramatically (Kavlak et al., 2017). According to National Renewable Energy Laboratory's (NREL) U.S. Solar Photovoltaic System Cost Benchmark, the price per watt of installed capacity fell from $7.34 in 2010 to $2.70 by 2018. These declining costs have made investing in a residential solar system more feasible for a greater portion of the population. In addition to declining costs, the prevalence of grid-connected residential solar energy programs have increased. Where available, solar energy produced in excess of demand by an individual's system can be "sold" back to the local electric utility. There exist different mechanisms by which individuals are compensated for their excess generation). This "sale" of excess generation creates the potential to produce long-run net-profits for those who install systems.

Aside from financial motivations, increasing concern over environmental issues has increased interest into environmentally conscious investments. This is reflected by the change in funding allocated by governments toward climate change related issues. In the US, annual investments of this nature by the federal government increased from $2.4 billion in 1993 to $13.2 billion in 2017 (Government Accountability Office, 2018). For context, the US federal government has allocated a total of $8.1 billion dollars to the National Science Foundation for fiscal year 2020 (Ambrose, 2019). Environmental advocates argue in favor of these investments, specifically those targeted at reducing greenhouse gas emissions, because of the theorized potential to slow the advance of climate change and avoid greater future

costs. Residential solar is one such investment. Residential solar can reduce overall demand for electricity from utilities and result in a reduction in net carbon emissions. Energy produced from a residential solar installation is carbon-free whereas a majority of grid generation is carbon intensive (63.5 percent of total US utility generation in 2018 was comprised of carbon-based fuels according to the U.S. Energy Information Administration (EIA)).

At a larger scale, governments have had an impact on the development of residential solar. From the federal government to local municipalities, various incentives have been devised to increase the number of residential installations. Yet, the motivation of governments to do so can be debated. It cannot be known whether politicians act out of genuine environmental concern or if renewable energy is simply a beneficial political position. Regardless, these incentives are aggressively pursued. A variety of incentives including property and sales tax rebates and cash discounts have been implemented. One such incentive is the Residential Renewable Energy Tax Credit (RRETC). Instituted in 2005, the RRETC allows taxpayers to claim a federal income tax credit equal to a percentage of expenditures on eligible renewable energy technologies (Bipartisan Budget Act of 2018). This is one example of many programs designed to spur growth in renewable energy technologies with a specific focus on residential solar.

These incentives are not free, however, and come at a significant cost to the tax base. This naturally led to contentious debate as to the efficacy and equity of such incentives. Advocates champion the potential environmental and grid benefits, while opponents of such measures argue that the benefits of these incentives disproportionately benefit wealthy individuals and that many of those who are installing systems would do so regardless of

incentives. This disproportionate benefit attained by the wealthy could represent a significant cross-subsidization. Cross-subsidization in this context is the subsidization of residential solar for wealthy individuals by middle- and low-income individuals. If this is the case, it presents an important equity concern to policy makers.

Previous research informed the debate by analyzing various solar incentives. Yet, this research is limited by the lack of a comprehensive database of residential solar installations in the United States. Some regions kept better records than others, but it remained a disjointed collection of data from differing regions. Institutions such as the Solar Energy Industries Administration (SEIA) and the Interstate Renewable Energy Council (IREC) attempted to use industry, utility, and state data to create more comprehensive databases. These resources are only available for purchase often at a significant cost. There have also been publicly available resources such as the Open PV Project by National Renewable Energy Laboratory (NREL) and Berkeley Lab's Tracking the Sun dataset. Both the public and private datasets have been used in research, but they share a common weakness; the presumed validity of these resources is only supported by approximately matching the estimated annual generation from residential solar.

The lack of a comprehensive resource left policy makers in the US to rely upon domestic research utilizing second best resources, or foreign studies, to guide domestic policy design. Yu et. al. (2018) attempts to solve this issue by developing the DeepSolar database. DeepSolar is a comprehensive database of residential solar installations in the United States based on satellite imagery. The theory behind the approach is that solar panels must have a clear view of the sky to function, so every residential solar installation must be visible with satellite imagery. To accomplish this, they developed a machine learning

algorithm to analyze and accurately identify residential solar installations from satellite imagery. Currently, the DeepSolar database only contains the number of systems installed by the date of the imagery analysis, but it is possible for this database to be continually updated in the future.

Much of the research to date on residential solar focuses on the impact of incentives on residential solar growth. One recent study by Matisoff & Johnson (2017) performed an analysis of 400 various state and utility incentives to quantify their respective impacts on residential solar growth. Of the incentives, some operate by reducing the initial cost burden of installation while others employ longer-term compensation strategies. They found incentives of the former resulted in the greatest impact while 67 percent of overall incentives had no statistically significant impact on residential solar.

Research has sought to create predictive solar development models. Some approaches focused on predictability alone at the cost of causal interpretation. Yu et. al. (2018) developed a Random Forest Regression model using the DeepSolar database. The Random Forest Regression is a machine learning process whereby a decision tree is formed by fitting a variable to randomly selected training and testing portions of the data. The model produced significantly better predictability and out-of-sample fit, but such a process produces a decision tree that has little causal interpretation. Other approaches have applied more traditional econometric methods with an emphasis on causal interpretation. De Groote et. al. (2016) analyzed residential solar data in Flanders, Belgium. They developed a series of Poisson regression models focused on capturing the heterogeneity in solar adoption.

There are two potential problems in the existing literature: the first is the reliance on incomplete databases, as discussed earlier, and the second is model selection. Many previous studies have relied upon Zero-Inflated Poisson and Zero-Inflated Negative Binomial

regressions. These modified forms of the Poisson and Negative Binomial Regressions attempt to account for data with an overabundance of zero observations, otherwise known as zero-inflated data. Zero-inflation is a likely condition of residential solar installation count data, as adoption has not become so widespread as to sufficiently reduce the prevalence of zero counts. Previous studies have selected these models in large part by relying upon the Voung Test, a likelihood-ratio test that compares the predicted probabilities of two non-nested models. Recent research has shown this test to have inherent bias in favor of the zero-inflated alternatives and to be unfit for determining the presence of zero inflation (Wilson, 2015).

I address these issues by applying the comprehensive DeepSolar database and a model selection process based on information criterion that includes analysis of Censored Regression Models. These models are specifically designed to handle data where there is evidence of censoring. I hypothesize that there is censoring in the zero-inflation process because it may be generated by some underlying censoring mechanism. That mechanism may be the theoretical zero lower bound of count data where an individual may have a negative propensity to install a residential solar installation, but their decision-making process is bounded by zero. If an individual finds residential solar to be unsightly or perceives residential solar as a threat they may have a negative disposition towards residential solar that would be reflected in a negative propensity to install. Thus, censored regression models which attempt to parameterize the censoring mechanism could perform better than a modified count distribution as applied in the Generalized Linear Model framework. Lastly, relying on information criterion rather than the Vuong Test ensures the selection process is unbiased.

I intend to provide insight into the demographic characteristics of residential solar adopters. By better understanding the population that is installing residential solar, researchers can attain a better understanding of the impacts of solar incentives. As previously mentioned, much of the debate around solar incentives revolves around the contention that cross-subsidies benefit wealthy residential solar adopters. By evaluating demographic characteristics such as mean income, income inequality, and the value of owner-occupied housing units, this study seeks to better identify the populations that have installed residential solar and gain a clearer picture of the efficacy and equity implications of solar development.

I find that a Cragg Hurdle Model specification outperforms the Generalized Linear Models applied in previous research. Based on the results of the model, I find evidence that state-level residential solar incentives have a positive impact on residential solar development. I also find that middle-class populations pursue residential solar development suggesting that a cross-subsidy for residential solar is not present.

CHAPTER 2: DESCRIPTION AND DATA

In this study, I use data to model, at the county level in the contiguous United States, the demographic characteristics of households that adopted residential solar. In the estimated models, the count of residential solar is the dependent variable and a selection of demographic variables are the independent variables. I regress the demographic variables on the count of residential solar installations to develop the economic relationship between the two. I describe the variables selected, the economic intuition for their inclusion in the model, and a description of the variable source. I use Stata 16 to conduct the econometric analysis.

I use the DeepSolar Database as the source of information on the count of residential solar installations in the US. Treating this variable as the dependent variable lets the developed models capture the impacts of various demographic variables on the development of residential solar. The DeepSolar database was created by developing a machine-learning algorithm to accurately identify installed residential solar from satellite imagery. Being the most comprehensive of the available datasets on residential solar installations, this resource enables a thorough analysis of residential solar development across the United States.

The remaining variables used in this model comprise the set of independent variables used to estimate changes in residential solar development. Table 1 contains descriptions and sources for all the variables used in this study. Table 2 contains the summary statistics for those variables.

Solar irradiance, measured in $kWh/m^2/day$, is a measure of the level of potential solar generation available to a location on the surface of a planet. I expect that residential solar development increases as the level of solar irradiance increases. This is because higher solar irradiance means greater potential solar generation, which allows a system to produce

better return on investment. I sourced this variable from the National Renewable Energy Laboratory (NREL).

Table 1: Variable Descriptions and Sources

Variable	Description
FIPS [1]	Five Digit FIPS Code
State [1]	State
County [1]	County
System Count [2]	Number of residential solar systems in 2017
Irradiance [3]	Solar radiation (kWh/m^2/day)
Population Density [4]	Thousand people per square mile (population/1000/mi^2)
GDP [5]	2017 real GDP in millions of 2012 chained dollars
RPP [5]	Regional Price Parity (RPP) by MSA
Democratic Vote [6]	Percent of presidential vote received by Democrats in 2016
Rural [7]	Dummy variable indicating a region is not part of an MSA
Housing Value [1]	Median normalized value of owner-occupied housing units ($/1000)
Owner Occupancy [1]	Percent of owner-occupied housing units
Mortgage Rate [1]	Percent of owner-occupied housing units with mortgage
Gini [1]	Estimated Gini coefficient (0 = perfect income equality)
Median Age [1]	Median age of the population
Unemployment Rate [1]	Unemployment rate (proportion of unemployed per 100 individuals)
Mean Income [1]	Mean household income in thousands of normalized dollars ($/1000)
State Incentives [8]	Number of state incentives for residential solar energy
Net Metering [8]	Number of years since the start of state net metering policy
Feed-in Tariffs [8]	Number of years since the start of state feed-in tariffs
Sales Tax Rebate [8]	Number of years since the start of state sales tax incentives
Retail Rate [9]	Average normalized residential retail rate of electricity from 2013 through 2017
Less Than High School [1]	Percent of population with less than a 12th grade education
High School Education [1]	Percent of population with high school education
Some College Education [1]	Percent of population with some college education
Associate Degree [1]	Percent of population with an associate degree
Bachelor's Degree [1]	Percent of population with a bachelor's degree
Graduate Degree [1]	Percent of population with a graduate degree

Note. Superscripts denote the data sources as follows:
(1) US Census Bureau's American Community Survey 5-Year Estimates 2017
(2) DeepSolar Database
(3) National Renewable Energy Laboratory
(4) US Census Bureau National Counties Gazetteer File
(5) Bureau of Economic Analysis – 2017 Data Tables
(6) MIT Election Data and Science Lab - County Presidential Election Returns 2000-2016
(7) National Bureau of Economic Research 2017 County Crosswalk File
(8) N.C. Clean Energy Technology Center's DSIRE Database
(9) Energy Information Administration

Table 2: Summary Statistics

Variable	Units	Mean.	Std. Dev.	Min	Max
System Count	-	447	3822	0	96874
Irradiance	kWh/m²/Day	4.98	0.502	3.72	6.75
Population Density	Population/1000/mi²	0.0006	0.00192	.000008	0.0963
GDP	$ Millions	5983	27969	17.8	688661
Regional Price Parity	-	0.908	0.0759	0.816	1.18
Democratic Vote	%	0.313	0.151	0.0314	0.909
Rural	Binary 1=non-MSA	0.631	0.483	0	1
Housing Value	$ Thousands	0.696	0.443	0.0009	1.77
Owner Occupancy	% x 100	71.6	7.98	19.7	93.5
Mortgage Rate	% x 100	51.9	12.3	0	85.9
Gini	-	0.445	0.0353	0.339	0.598
Median Age	-	41.2	5.31	23.5	66.4
Unemployment Rate	% x 100	6.33	2.98	0	28.7
Mean Income	$ Thousands	59.7	18.7	25.7	187
State Incentives	-	5.76	3.7	1	17
Net Metering	-	12.3	9.27	0	35
Feed-in Tariffs	-	0.351	1.81	0	11
Sales Tax Rebate	-	3.12	6.3	0	40
Retail Rate	Cents	8.66	2.23	6.32	18.04
Less Than High School	Exclusion Category	13.8	6.5	1.1	58.6
High School Education	%	34.5	7.1	8	54.9
Some College Education	%	21.8	3.81	6.5	36.6
Associate Degree	%	8.74	2.612	0.8	21.7
Bachelor's Degree	%	13.7	5.54	2.4	43.7
Graduate Degree	%	7.31	4.12	0	40.3

I also include Gross Domestic Product (GDP) and Regional Price Parity (RPP) in this analysis. GDP represents the total economic output at the county level and is expected to show a positive correlation with residential solar development. As a region's GDP grows, the region's overall wealth increases which is expected to increase the development of residential solar within the region. RPPs represent price level adjustments across the United States. Purchasing power across the United States is not homogenous, so it is necessary to use RPP to normalize other pricing variables included in this study to ensure that results are not skewed by the price level. I retrieved both variables from the Bureau of Economic Analysis (BEA).

I include population density and a dummy variable for Metropolitan Statistical Area (MSA) to capture the effects of urbanization on residential solar development. Population

density, measured as thousand people per square mile, yields a parabolic relationship with residential solar development. I expect that residential solar development will be lowest in areas with the highest population densities, which correspond to urban environments where housing units are primarily condos and apartments. The parabolic shape is likely to result from an optimality in increasing population density corresponding to a point where the housing characteristics turn from suburban areas to more urban regions. Inclusion in an MSA captures the effect of urban economies as compared to rural economies. An MSA is a region of interrelated economies that transcends localized borders. Inclusion of the MSA variable captures the effect of proximity to economic centers and is expected to have a positive impact on residential solar development, as a more robust economy should generate greater wealth. I sourced data on population density from the US Census Bureau National Counties Gazetteer File. I use information from the National Bureau of Economic Research (NBER) on Metropolitan Statistical Areas.

The average residential retail rate of electricity per kWh included in these models captures the effect of the cost of electricity on residential solar development. Intuitively, I expect residential solar development to increase as the price of electricity increases since this would make the return on investment of a residential installation greater. I normalize these values with the RPP. I sourced data on the per kWh retail rate of electricity from the EIA.

I expect that personal political ideology has an impact on an individual's propensity to adopt residential solar. To capture this effect, I use the percentage of votes cast for the Democratic candidate in each county in the 2016 Presidential Election . I theorize that, on average, Democratic voters tend to have a more favorable view of renewable energy technologies and a greater willingness to invest in environmental measures. This would cause

the level of residential solar development to increase as the percentage of Democratic votes increases. Election results are from the MIT Election Data and Science Lab.

I include data on the number of state solar incentives, the number of years since the implementation of net metering, the number of years since the implementation of feed-in tariffs, and the number of years since the implementation of a sales tax rebate program to capture the effects of solar incentives on the development of residential solar. Since these are all state incentives designed to stimulate the growth of residential solar development, I expect that all of these would have a positive impact if they are effective. I use the N.C. Clean Energy Technology Center's DSIRE Database as the source for information on solar incentives.

The number of state solar incentives is a proxy for the state level investment into solar. Ideally, information on each incentive and its value could be incorporated to gain more accurate insights, but the exact implementation of incentives and the legislative structure of such incentives vary greatly. This simplification allows for a general analysis of the impact of increasing solar incentives.

The number of years since the start of net metering captures the impact of net metering policies. Net metering is a compensation framework for residential solar wherein the system owner receives energy credits for unused solar generation transmitted to the connected electric grid. The number of years since the start of feed-in tariffs captures the impacts of state feed-in tariff policies. Feed-in tariffs guarantee solar adopters fixed value for their solar generation for a period of years. In theory, this should encourage solar development by providing long term certain returns to solar adopters. Lastly, I include the number of years since the start of a sales tax rebate program to capture the effect of sales tax rebate programs on residential solar development. These programs exempt residential solar

installations from sales tax on purchased components. This lowers the upfront cost of such a system and likely incentivizes the development of residential solar.

The remaining demographic variables are all from the US Census Bureau's American Community Survey 2017. This survey provides publicly available information on a wide variety of topics across the United States. This study utilizes the five-year estimate version of the dataset. This version of the survey includes the most granular data and provides the best insight into the true population values by averaging the estimates across the five-year period preceding the date of publication.

The median normalized value of owner-occupied housing units captures the value of housing within regions having residential solar. If solar incentives are encouraging the wealthy to install residential solar, then we would expect a positive relationship between the value of housing and residential solar development. I use only data on owner-occupied housing because renters are not likely to invest in residential solar. A tenant has little to gain from investing in a residential solar installation as the return on investment spans decades . A landlord similarly has little incentive to install systems on their rental properties because tenants most often pay utilities which would negate landlord benefits of installing a solar system. For this same reason, I include the percentage of owner-occupied housing units in each county.

The last of the housing characteristics included in this analysis is the percentage of households with mortgages. Since a residential solar installation presents a significant investment, which may require a loan to purchase for most households , the level of housing debt an individual has may impact the rate of solar development. Mortgages, representing home debt, may be a significant consideration for homeowners when evaluating residential solar. The return on installing a residential system may not outweigh the cost particularly

when the return is weighted by existing debt. This means that residential solar is likely to have lower development in regions with greater proportions of homes with mortgages.

Educational attainment of the population and median age are also included. Median age is intended to capture generational differences in solar adoption. Younger generations are likely to be more environmentally conscious and thus more likely to invest in a residential solar system. More educated individuals are likely to have a greater propensity to support environmental issues, so the number of residential solar installations is likely to increase as educational attainment increases. I consider education at several levels: less than high school, high school degree, some college education, associate's degree, bachelor's degree, and graduate degree. These variables contain the percentage of the population in each county by their highest level of educational attainment. The category of less than high school is the exclusion category in each estimated model.

The remaining variables are the Gini coefficient, the unemployment rate, and mean income. The Gini coefficient measures the level of income inequality in a region. A Gini coefficient of 0 indicates perfect income equality and a coefficient of 1 indicates perfect income equality. If only the wealthiest individuals within a region pursue residential solar , then there will be a positive correlation between the Gini coefficient and the count of residential solar. The unemployment rate is also included to help describe the characteristics of regions that have adopted residential solar. If residential solar is pursued in areas characterized by higher levels of unemployment, then the unemployment rate will have a positive correlation with residential solar development. The last of these variables is mean income. Inclusion of mean income, normalized by the Regional Price Parity, gets at the core of the cross-subsidization issue. If cross-subsidization is taking place, then we expect to see a positive correlation between mean income and residential solar development.

CHAPTER 3: MODEL SELECTION

3.1 Dependent Variable Characteristics

The variable of interest in this study is the count of residential solar installations in each county of the contiguous United States. In general, the distribution of a count variable will have significant impacts on the methods selected to model its behavior. Intuition suggests that these data contain clustering at the zero-lower bound and characteristics of zero inflation driven by censoring. Zero inflation is a condition wherein the data process is characterized by an overabundance of zero observations. Clustering is a condition wherein a significant portion of observations are contained within a small portion of the distribution relative to the size of the sample. Clustering in this application is likely a result of the fact that residential solar is not prevalent in most of the United States, so the number of installed systems in most counties is relatively low. The censoring occurs because the propensity to install a residential system is bounded by zero, even though propensity could be negative. A negative propensity to install captures the effects of individuals who possess a negative disposition towards residential solar resulting from various reasons such as personal preference or hostile political positions. This censoring leads to an inflated number of zero observations.

It is important to note the difference between censoring and truncation. These two definitions are often associated with count data models and can lead to confusion. Truncation occurs when values that exceed a threshold are dropped from the data. Censoring, on the other hand, is a process by which values that exceed a threshold are contained within a value. This can be visualized as an inequality operator where in this instance zero is not just zero but a placeholder that represents all values less than or equal to

zero. Simply put, truncation excludes the additional information and censoring includes it, albeit inside a different value.

I first visually inspect the data for evidence of zero-inflation and censoring. Figure 1 contains a histogram of the number of residential solar installations in the US. This figure provides graphical evidence of censoring at the zero-lower bound, as well as significant clustering of values near zero.

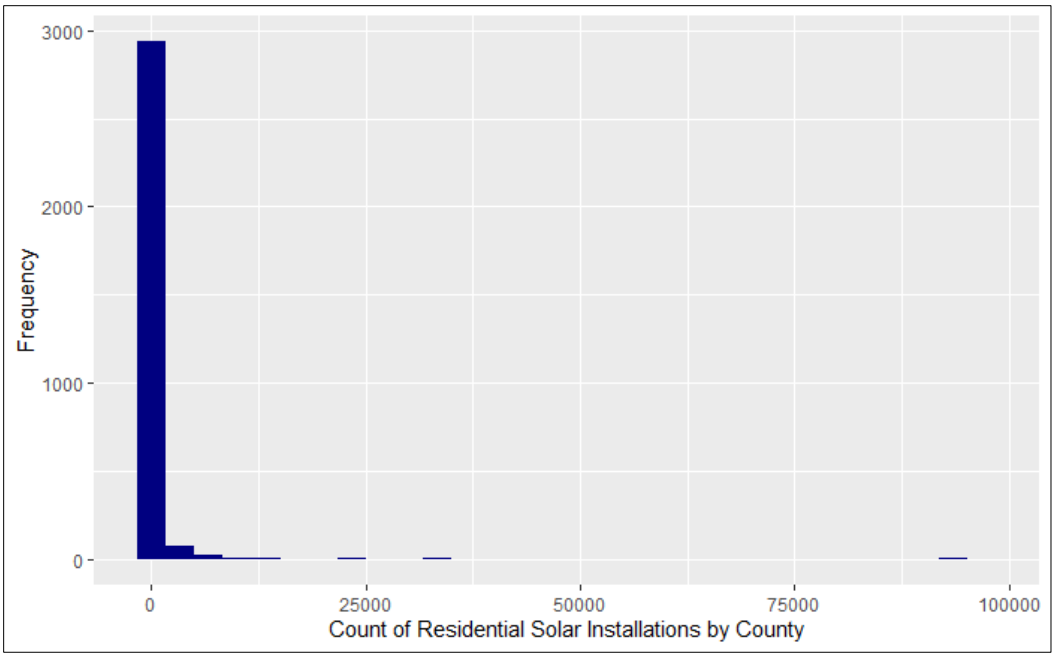

Figure 1. Histogram of the number of residential solar installations in the US

This histogram is limited in its ability to provide further insight into the remaining distribution of count values because it is heavily driven by lower count observations. Examining the summary statistics and variable distribution can assist in this evaluation. The number of zero observations constitutes approximately 12 percent of the entire sample while approximately 56 percent of the observations contain ten or fewer systems. With a mean count of 448, a maximum observation of 96,874, and a standard deviation of 3,823, the

distribution is certainly clustered near zero and contains a sizeable number of zero observations. This is not to suggest that the mere presence of many zeros is evidence enough to imply zero inflation. Zero inflation is a specific instance that occurs when the underlying process by which those many zeros are present is driven by a different underlying process than the rest of the observations. I will discuss this further in the latter portion of this chapter.

Having established with theory and evidence that the count-based dependent variable may be censored at the zero-lower bound, it is important to contemplate bias in the data generation process. If the process by which count data is collected is biased, then the data itself contain bias. In this instance, there is no reason to expect that any values produced by the DeepSolar method were altered in an intentionally systematic way. In fact, with the high level of accuracy achieved by the methodology, it may be safe to assume that the exclusions are minimal (Yu et. al., 2018) and likely to be the result of randomness rather than systematic bias.

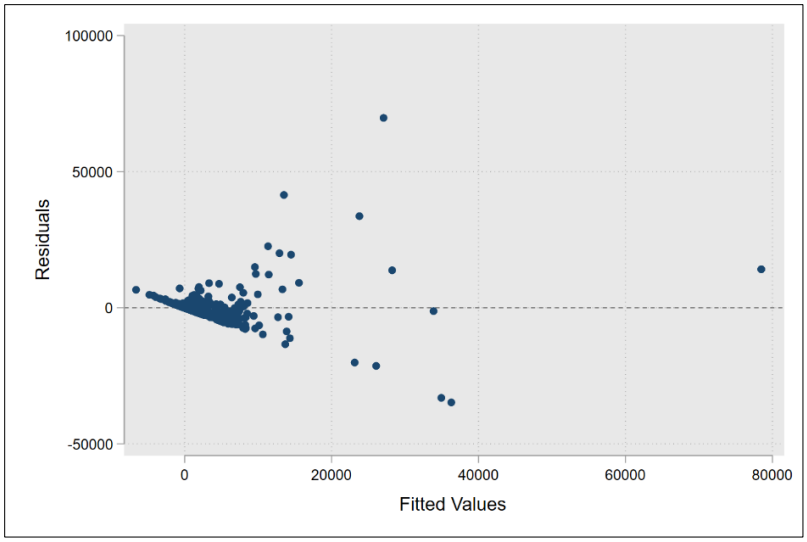

Figure 2. Residuals vs fitted values of the OLS regression

In the presence of censoring and zero-inflated data, the normality assumption is likely violated. As an example, a simple Ordinary Least Squares (OLS) regression was fit on the full model. In the presence of censored data, standard OLS will be inconsistent, but the residuals can still provide insight. Figure 2 contains a plot of fitted values versus residuals. This plot shows a clear picture of heteroskedasticity in the error. Relying on normality as the underlying distributional assumption is not likely a valid approach for this data.

3.2 Generalized Linear Models (GLM)

3.2.1 Generalized Linear Model Framework

A censored count variable as the dependent variable suggests that a Generalized Linear Model (GLM) is appropriate. GLMs are of the family of Maximum Likelihood Estimators (MLE) that seek to maximize the log-likelihood of the general form shown in Equation 1:

$$Q(\theta) = \sum_{i=1}^{N} \ln f\,(y_i|x_i\,,\theta) \quad , \tag{1}$$

where $f(y_i|x_i,\theta)$ is the conditional density for a continuous y (Cameron and Trivedi, 2009). Generalized Linear Models allow for the underlying probability distribution of the dependent variable to differ from the standard normal distribution as applied in Ordinary Least Squares regressions. This is done by associating the response variable to the linear model by a link function based on a linear exponential function. The general form of the GLM estimator is shown in Equation 2:

$$Q(\theta) = \sum_{i=1}^{N} [a(m\,(x_i,\beta)) + b(y_i) + c(m(x_i,\beta))y_i] \quad , \tag{2}$$

where $m(x_i, \beta) = E(y|x)$; $E(y|x)$ is the conditional mean of y, a(-) and c(-) correspond to different exponential linear functions, and b(-) is a normalizing constant (Cameron and Trivedi, 2009).

The count data perspective and the censoring of the data are two separate issues to be addressed by the GLM framework. GLMs in general are designed to better encapsulate count data behavior by allowing the count distribution to vary with the link function. Special versions of these GLMs are specifically designed to address the censoring. A zero-inflated specification treats the outcome variable as a two-part process. This allows zeros in the count data to be the result of some binary function and a predictive count function (Cameron and Trivedi, 2009).

As discussed previously, the censoring in this instance likely results in zero inflation, so a specification that accounts for the zeros would provide a better alternative. It is possible that the zeros do not represent inflation in which case the nonzero-inflated version of the GLM will perform better. In this study, I apply the Poisson and the Negative Binomial distribution as alternative linear exponential link functions.

3.2.2 Poisson Regression Model (PRM)

The first of the probability distributions to be evaluated is the Poisson distribution. This is often the first distribution applied in the presence of count data. The Poisson regression is a subset of GLM that assumes that the response variable follows a Poisson distribution of the form in Equation 3:

$$f(y|x) = e^{-\mu}\mu^{y}/\mathrm{y}! \quad , \tag{3}$$

where μ is the conditional mean given by $E(y|x) = \exp(x'\beta)$. The exponential form of the conditional mean ensures that it is positive and in line with non-negative count values and

thus the solutions to the maximum likelihood process produce estimates where the count occurrence is both positive and dependent upon a Poisson distribution (Cameron & Trivedi, 2009).

The PRM is a parsimonious approach compared to other GLMs, but it has two restrictive assumptions. First, it assumes the independence of events. This is a particularly troublesome assumption in this study as it is likely that the installation of solar accelerates as a more robust residential solar community develops in a region. Second, the PRM assumes that the data exhibits equidispersion, meaning, the conditional mean is equal to the conditional variance. This condition is commonly violated in the presence of count data and requires estimation with robust standard errors or a different distributional assumption (Cameron & Trivedi, 1986).

3.2.3 Negative Binomial Regression Model (NBRM)

The Negative Binomial Regression Model is a GLM where the underlying link function is based on the negative binomial distribution. The NBRM is a generalized form of the Poisson Regression Model that relaxes the condition of equidispersion. The application of the NBRM used in this study, known as NB2, is built on a Poisson-gamma distribution. This gamma distribution is shown in Equation 5:

$$f(\lambda_i) = \frac{1}{\Gamma(v_i)}\left(\frac{v_i\lambda_i}{\phi_i}\right)^{v_i} \exp\left(\frac{-v_i\lambda_i}{\phi_i}\right)\frac{1}{\lambda_i} \quad , \tag{5}$$

where $\lambda_i \sim \text{Gamma}(\phi_i, v_i)$, ϕ_i is the mean, v_i is the smoothing parameter, $\lambda_i > 0$, $\phi_i > 0$, and $v_i > 0$. From this distribution, the variance is estimated as shown in Equation 6:

$$Var(y) = \phi_i + \alpha\,\phi_i^2 \quad , \tag{6}$$

where α is the adjusted smoothing parameter of the form $(1/v_i)$ from Equation 5. This allows the variance to be greater than the mean and thus allows for inter-regional heterogeneity (Cameron & Trivedi, 1986). The smoothing parameter α is used to minimize the interference of noise in the probabilistic process. This term is parameterized in the estimation process and as it approaches zero, the NBRM converges to the PRM.

3.2.4 Zero-Inflated GLM Framework

A weakness of standard models in the GLM framework is their inability to properly capture the process of zero inflation. This led to the development of Zero-Inflated GLMs whereby the zeros are modeled by a more flexible process involving the count density and a binary process. Zeros can result in two ways: when the binary process is equal to zero or when the binary process is 1 and the count density then equals zero. This functional relationship is modeled in Equation 7:

$$f(y) = \begin{cases} f_1(0) + \{1 - f_1(0)\} f_2(0) & if\ y = 0 \\ \{1 - f_1(0)\} f_2(y) & if\ y \geq 1 \end{cases}, \quad (7)$$

where $f_2(*)$ represents the count density and $f_1(*)$ is the density function of a binary process. If the binary process equals 0, then y = 0, with a probability of $f_1(*)$. If the binary process equals 1, then y draws a value from the count density $f_2(*)$ with a probability of $f_1(1)$. This flexible framework allows the parameters that describe the two densities to differ. This approach can be applied to both the Poisson Regression Model (creating the Zero-Inflated Poisson Regression Model) and the Negative Binomial Regression Model (creating the Zero-Inflated Negative Binomial Regression Model) where $f_2(*)$ then represents the respective count density (Cameron & Trivedi, 2009).

3.3 Censored Regression Models (CRM)

3.3.1 Tobit Regression Model

GLMs are not the only available approach. Research on modeling with censored data has created its own category of censored regression approaches. The first and most basic of these models is the Tobit Regression Model. It attempts to account for censoring by modeling the outcome variable as a function of some unobserved latent variable. The issue in this instance is that the Tobit model relies heavily on normal residuals to produce consistent estimators and this characteristic is not likely to be present in these data. In a general approach, the data may be transformed to produce normal residuals. This analysis, however, relies on zero-inflated count data making it unwise to transform the data in traditional ways, i.e. log transform the dependent variable. Transformation of this nature would exclude the zero observations and give inaccurate results. Instead, using the original count data in an alternative censored model approach is the best course of action.

3.3.2 Cragg Hurdle Model

The alternative approach applied here is a specific form of hurdle model. Hurdle models, called two-part models, allow for the zero values in the dependent variable to be produced by two independent processes. They are designed to capture the two-step decision- making process in the dependent variable. First there is the decision to be in the participant group and then there is the decision to participate (Cameron and Trivedi, 2009). In terms of this study, the first model predicts whether any residential solar systems will be installed in a county and a second model predicts how many systems will be installed given the first model predicts that county will have solar. These two decisions are treated as two independent models and separately estimated.

I apply a specific version of Hurdle Model that is known as the Exponential Cragg Hurdle Model. This model assumes a lower bound of 0 and is characterized by the model in Equation 8:

$$y_i = s_i h_i^* \quad , \tag{8}$$

where y_i represents the observed value of the dependent variable and h_i^* is a continuous latent variable observed when the selection variable s_i is equal to 1. The term h_i^* in the exponential framework is of the form shown in Equation 9:

$$h_i^* = \exp(x_i\beta + v_i) \quad , \tag{9}$$

where x_i is a vector of explanatory variables, β is a vector of coefficients, and v_i is an error term with a normal distribution. The selection variable s_i is modeled in Equation 10:

$$s_i = \begin{cases} 1 & if\ z_i\gamma + e_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad , \tag{10}$$

where z_i is a vector of explanatory variables, γ is a vector of coefficients, and e_i is a standard normal error term (StataCorp, 2019a).

The specification used in this study applies a Probit approach to the selection variable, and then fits the exponential outcome model on the dependent variable. This process treats the values at zero (which is the censoring point in these data) as being observed rather than censored. This means that observations at the censored value are not treated as the result of an inability to observe the distribution below the censoring point (StataCorp, 2019a).

3.3.3 Heckman Two-step Regression Model

The Heckman Two-step Regression Model is another censored regression model commonly used in data analysis. The approach is similar in nature to the Cragg Hurdle Model in that it attempts to account for a sample selection issue by first modeling the sample selection via a Probit specification and then modeling the behavior with least squares regression. However, the Cragg Hurdle is designed to account for censorship in the data and the Heckman Two-step Model is designed to account for truncation in the data. Given that the dataset of interest in this study is assumed to be censored and not truncated, the Cragg Hurdle approach is more appropriate (Cameron and Trivedi, 2009).

3.3.4 Stochastic Frontier Analysis

The last of the relevant censored regression model approaches is Stochastic Frontier Analysis. Originally designed to model inefficiency in the production of a firm, stochastic frontier models have expanded to analyze a variety of censored econometric problems. They have also been used to analyze consumer demand. By modeling the deviation from the estimated frontier, researchers can estimate the demand characteristics of a sample. The general approach is to model the outcome error as some function of an efficiency term and then log transform both sides of the equation. This approach is not applicable in this instance due to the use of count level data with likely zero inflation. A transformation of the data as applied in Stochastic Frontier Analysis will alter the data in such a way as to distort the model results. For this reason, this model is not pursued further in this study (StataCorp, 2019a).

3.4 Overdispersion Testing

The first of the Generalized Linear Models to be estimated is traditionally the parsimonious Poisson regression. If the Poisson is properly specified and fits the data well, it

can outperform the more complicated alternatives. The first step in evaluating the Poisson regression is to verify the property of equidispersion. In a properly specified Poisson model, we expect to find that the conditional variance is equal to the conditional mean. Often this property is violated, in which case there is overdispersion in the data and additional steps must be taken. Table 3 contains the results of the Deviance and Pearson goodness-of-fit tests for the Poisson regression.

Table 3: Overdispersion Tests for the PRM

Deviance goodness-of-fit:	929961
Prob > χ^2 (df = 3024)	0.000
Pearson goodness-of-fit:	1531565
Prob > χ^2 (df = 3024)	0.000

Note. Null hypothesis is the presence of equidispersion.

These tests measure how well the Poisson model fits the data. With p-values of zero, both reject the null hypothesis of equidispersion. This suggests that a Negative Binomial regression or a Poisson regression with robust standard errors may provide more accurate estimators of the probabilistic outcomes than the standard Poisson regression.

In all, I estimate a Poisson regression (PRM), a Negative Binomial regression (NBRM), a Zero-Inflated Negative Binomial regression (ZINB), and a Zero-Inflated Negative Poisson regression (ZIP) on the full model with an inflation specification containing all variables. All models are estimated using robust standard errors. Figure 3 and Table 4 display the dispersion values for each model. All the models struggle to properly model the zero counts to some extent. From the graph, the PRM presents high levels of overdispersion relative to the other models which exhibit underdispersion. This suggests that a NBRM will outperform the PRM due to its ability to parameterize the overdispersion.

However, it is likely that the zero-inflated alternatives will outperform the NBRM in the presence of underdispersion because the NBRM has difficultly estimating the additional smoothing parameter under these conditions. Of the zero-inflated alternatives, the ZINB presents the greatest dispersion, but this doesn't automatically imply that the ZIP is the preferred model. It does, however, raise the possibility that the underdispersion could present difficulties in properly modeling the ZINB. The numerical approximation methods of the ZINB tend to have difficulties in the presence of underdispersion.

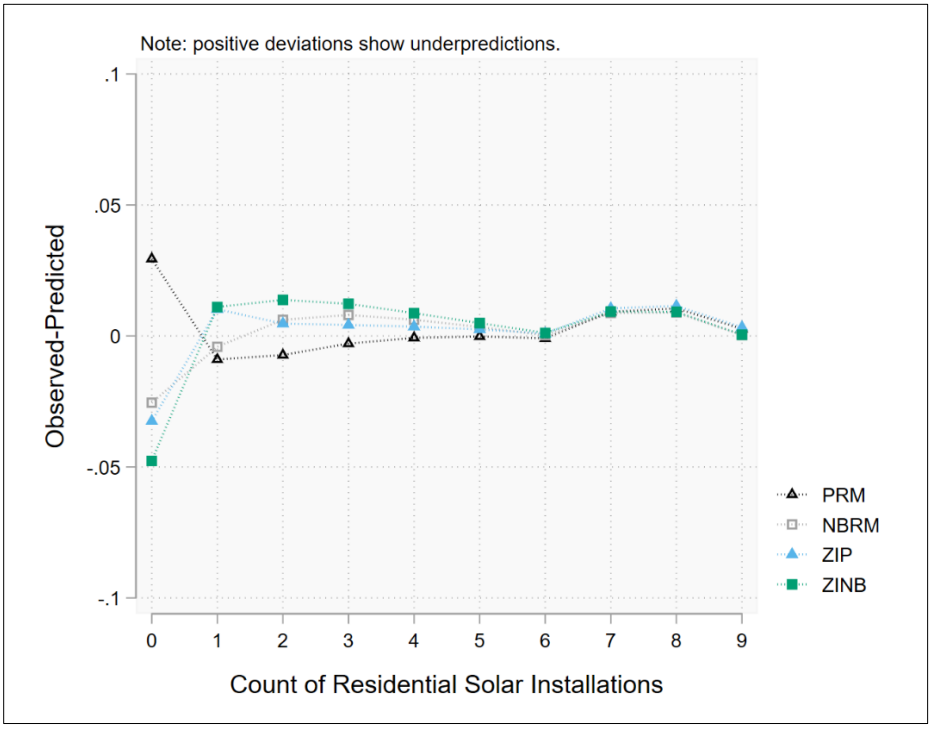

Figure 3. Dispersion Comparison of the GLM Methods

Table 4: Dispersion Values for the GLM Methods

Model	Maximum Difference	At Value	Mean Difference
PRM	0.029	0	0.007
NBRM	-0.025	0	0.007
ZIP	-0.033	0	0.008
ZINB	-0.048	0	0.012

3.5 Model Performance

A comparison of model performance across all the estimated models is contained in Table 5. All models had the same specification. The dependent variable was system count. The independent variable set was comprised of the study dataset excluding educational attainment, real GDP, and population density. I selected these independent variables for the sake of numerical optimization and parsimony among all the models.

Table 5: Performance Statistics of GLMs and CRMs

Model	Log-Likelihood Ratio (null)	Log-Likelihood Ratio (model)	df	AIC	BIC
PRM	-4503954	-619938	20	1239916	1240036
ZIP	-4330739	-611359	38	1222794	1223023
NBRM	-15478.6	-12734.6	21	25511.26	25637.79
ZINB	-15433.4	-12691.6	39	25461.13	25696.11
CRAGG	**-12017.1**	**-10226.3**	39	**20530.67**	**20765.66**

Note. Bold values represent the best performance. df: Degrees of Freedom. AIC: Akaike Information Criteria. BIC: Bayesian Information Criteria. Dependent variable was the count of residential installations. The independent variable set was comprised of the study dataset excluding housing built and education percentages. Selector variable set was the same as the independent variable set excluding GDP and population density. These selections were made for the sake of numerical optimization and parsimony among all the models.

Based on the log-likelihood ratios and both the AIC and BIC statistics, there is strong evidence that the Cragg Hurdle Model is the preferred model overall. While the two-step modeling of the dependent variable presents a more complicated model than either the PRM or the NBRM, the causal interpretation of the Cragg Hurdle model is more straightforward than that of the ZIP or the ZINB. Based on these results, the Cragg Hurdle Model is the best approach.

CHAPTER 4: RESULTS

The Cragg Hurdle specification applied in this study is a two-step approach employing a first stage Probit Regression and a second stage Zero Truncated Negative Binomial Regression. In terms of this study, the first stage models the likelihood of a county looking to install residential solar and the second stage models the amount of residential solar installed given that a county has at least one homeowner who has chosen to install solar. A unique feature of the two-step approach is the ability to independently evaluate the first stage of the model before combining it with the second stage of the overall model. This allows the selection decision as modeled by the first stage to be analyzed further. The results of the Cragg Hurdle Model are contained in Table 6. The independent variables included are assumed to be exogenous to the process such that the model results are interpreted to be causal.

The coefficients produced by the Cragg Hurdle, in both the first and second stage, cannot be interpreted in the traditional linear sense. These coefficients represent the log odds ratios which is an alternative specification for probability. Odds ratios represent a proportion of theoretical outcome and the log of the odds ratio is often applied in models to overcome issues that may arise with small sample data. The sign of coefficients can still be analyzed for a directional impact in the first stage selection model. The variables that are significant can provide some context to the second stage model, but neither the sign nor magnitude provide insight as its effects are reliant upon the first stage.

Marginal effects of the overall model must be estimated to interpret the coefficients in a traditional sense. Marginal effects are calculated as the partial derivative of the independent variable with respect to the dependent variable. The average marginal effects

that evaluate the average change in predicted values for a change in the independent variable are used in this study.

Table 6: Cragg Hurdle Model Results

	First Stage			Second Stage		
Variable	Coef	Std.Err.	p-value	Coef	Std.Err.	p-value
Irradiance	-0.918	1.286	0.476'''	0.6073	0.051	0.000***
Irradiance ^{2}	0.124	0.124	0.317'''			
Gini	4.836	17.968	0.788'''	9.942	2.505	0.000***
Gini ^{2}	-1.285	19.833	0.948'''			
ln(GDP)	0.255	0.649	0.695'''	1.801	0.168	0.000***
ln(GDP) ^{2}	0.054	0.060	0.363'''			
Gini x ln(GDP)				-1.423	0.362	0.000***
Population Density	634.239	1366.849	0.643'''	193.879	108.237	0.073*
Population Density ^{2}	-2623.008	18594.470	0.888'''			
Population Density x ln(GDP)	-31.042	236.744	0.896'''	-17.383	8.112	0.032**
Rural				1.994	0.224	0.000***
Rural x ln(GDP)				-0.334	0.030	0.000***
Housing Value				0.557	0.187	0.003***
Housing Value ^{2}				-0.423	0.135	0.002***
Owner Occupancy	0.188	0.069	0.007***	0.149	0.024	0.000***
Owner Occupancy ^{2}	-0.001	0.000	0.005***	-0.001	0.000	0.000***
Mortgage Rate	0.065	0.023	0.005***	0.085	0.013	0.000***
Mortgage Rate ^{2}	-0.001	0.000	0.055**	-0.001	0.000	0.000***
Median Age	0.115	0.089	0.198	-0.193	0.052	0.000***
Median Age ^{2}	-0.002	0.001	0.091*	0.002	0.001	0.000***
Unemployment Rate				0.104	0.023	0.000***
Unemployment Rate ^{2}				-0.004	0.001	0.001***
Mean Income	-0.019	0.005	0.000***	-0.016	0.006	0.006***
Mean Income ^{2}				0.0001	0.000	0.005***
Democratic Vote				-0.491	0.255	0.054**
Net Metering				-0.006	0.006	0.572
Net Metering x Retail Rate				.004	0.001	0.004***
Democratic Vote x Net Metering				-0.077	0.016	0.000***
State Incentives				-0.051	0.017	0.003***
Democratic Vote x State Incentives				0.269	0.047	0.000***
Feed-in Tariffs				0.025	0.040	0.536'''
Democratic Vote x Feed-in Tariffs				0.107	0.084	0.205'''
Sales Tax Rebate				0.029	0.010	0.003***
Democratic Vote x Sales Tax Rebate				-0.005	0.020	0.802'''
Retail Rate				-0.043	0.027	0.111
High School Education	0.034	0.011	0.003***	0.017	0.006	0.013**
Some College Education	0.048	0.013	0.000***	0.019	0.007	0.004***
Associate Degree	0.070	0.018	0.000***	0.015	0.009	0.125'''
Bachelor's Degree	0.040	0.015	0.009***	0.023	0.008	0.007***
Graduate Degree	0.050	0.023	0.032**	-0.002	0.009	0.683'''

Note. The first stage represents the Probit selection model only. Second stage represents the Zero Truncated Negative Binomial where Coef is the coefficient as the log odds ratio and Std. Err is the robust standard error. *** p<0.01, ** p<0.05, * p<0.1, ''' joint significance

4.1 First Stage Selection Model

For many of the variables in the selection model, quadratic specifications are optimal. Gini, population density, mortgage rate, owner occupancy, and median age all have a parabolic specification where the quadratic terms imply an optimality point towards the middle of the distribution of the variable's values with a decreasing impact at the ends of the distribution. Irradiance and the log of GDP both show traditional quadratic relationships where the effect increases in the positive direction. This relationship implies that as the log of GDP or irradiance increases, the probability of being a solar county increases. All the coefficients on education were statistically significant and positive. This implies that increases to all levels of education have a positive impact on the probability of being a solar county as compared to the percentage of the population that has less than a high school degree.

4.2 Second Stage Model

In the second stage model, interactions between the percentage of Democratic vote and state incentives, sales tax rebate, and feed-in tariffs show that there is a statistically significant relationship. This implies that there are greater impacts of these incentives in regions with a higher percentage of Democratic voters. This would be expected as policies pursued by politicians are largely based on their constituency and any policies pursued are likely to have an electorate that would respond positively.

4.3 Complete Model

Table 7 contains the estimated average marginal effects for the full Cragg Hurdle Model. From these results, I derive a clear picture of the demographic characteristics of residential solar adoption. On average, residential solar installations are associated with populations that are younger, more educated, and lean Democratic politically. These results are in line with an intuitive assessment of residential solar, which would suggest that these populations are more willing to adopt such technologies.

Table 7: Average Marginal Effects of the Cragg Hurdle Model.

Variable	dy/dx	Robust Std. Err.	z-stat	p-value	95% Confidence Interval	
Irradiance	272.609	36.878	7.39	0.000***	200.3301	344.888
Gini	-2897.197	838.522	-3.46	0.001***	-4540.67	-1253.725
ln(GDP)	486.350	49.367	9.85	0.000***	389.5923	583.1069
Population Density	-2663.634	7448.877	-0.36	0.721	-17263.16	11935.9
Democratic Vote	907.255	266.128	3.41	0.001***	385.6542	1428.856
Rural	-369.865	37.124	-9.96	0.000***	-442.6257	-297.104
Housing Value	-176.131	60.351	-2.92	0.004***	-294.4175	-57.84461
Owner Occupancy	10.514	2.539	4.14	0.000***	5.537774	15.48952
Mortgage Rate	-4.370	2.800	-1.56	0.119	-9.858645	1.118617
Median Age	-6.553	2.926	-2.24	0.025**	-12.28756	-0.8191023
Unemployment Rate	20.560	4.523	4.55	0.000***	11.69576	29.42446
Mean Income	0.744	1.127	0.66	0.509	-1.465163	2.954087
Net Metering	0.989	3.938	0.25	0.802	-6.72873	8.706216
State Incentives	50.172	8.842	5.67	0.000***	32.84202	67.50167
Feed-in Tariffs	39.943	9.689	4.12	0.000***	20.95372	58.93203
Sales Tax Rebate	11.537	2.378	4.85	0.000***	6.877134	16.19732
Retail Rate	11.643	6.982	1.67	0.095*	-2.041334	25.32722
High School Education	7.550	2.748	2.75	0.006**	2.163323	12.93673
Some College Education	8.722	2.951	2.96	0.003***	2.938557	14.50564
Associate Degree	6.878	4.177	1.65	0.100*	-1.308823	15.06428
Bachelor's degree	10.298	3.514	2.93	0.003***	3.409836	17.1861
Graduate Degree	-0.644	3.852	-0.17	0.867	-8.192549	6.905343

Note. dy/dx is the average marginal effect. All errors are robust where: *** p<0.01, ** p<0.05, * p<0.1.

Regions with higher prevalence of residential solar have more solar irradiance, higher retail electricity rates, and greater economic output. These results are to be expected, as more solar irradiance suggests the ability to generate more energy annually, higher retail electricity rates suggest that installing solar has a shorter return on investment, and a larger economy suggests more overall wealth.

Regions with higher prevalence of residential solar are also characterized by less income inequality. The Gini coefficient, bounded by 0 and 1 with a value of 1 meaning there is perfect income inequality, represents a fractional change in income inequality. The negative average marginal effect of Gini implies that as the level of income inequality increases, the number of residential solar installations decreases. This suggests that income within a county that has more residential solar is more evenly distributed.

Household characteristics of regions with greater solar penetration have higher rates of owner occupancy and lower values of owner-occupied housing units. State residential solar incentives, on average, increase the number of residential solar installations by fifty per county for each additional incentive. Feed-in tariffs and sales tax rebates have a positive impact on the number of residential solar installations while mean income, population density, and net metering have a neutral average impact on the prevalence of residential solar.

CHAPTER 5: CONCLUSIONS

The results of this study present several contributions to the literature including the relevance of hurdle model approaches in modeling residential solar adoption, evidence to support the efficacy of solar incentives, and evidence against the presence of income cross-subsidization for residential solar.

This study highlights the need to evaluate hurdle models as a potential modeling mechanism when evaluating residential solar. They not only present a more intuitive econometric interpretation, but these results showed that a properly specified hurdle model approach (a Cragg Hurdle Model was most appropriate in this study) can outperform the more often applied Generalized Linear Models. The flexibility in the design of hurdle models may allow researchers to more accurately model the solar adoption process.

I find evidence to support the efficacy of solar incentives in increasing residential solar. The variable State Incentives, representing the total of all residential solar incentives, had a s positive and statistically significant average marginal effect. This would suggest that for each additional state solar incentive, the number of residential solar installations in each county will increase. It is possible with better data, that some incentives are the drivers of this result while others have a negligible impact. Of the two incentives that were modeled separately, it was found that feed-in tariffs and sales tax rebates both have a positive impact on residential solar development.

It was also found that net metering policies, on average, have an insignificant impact on the number of residential solar installations once political sentiment and the retail rate of electricity are considered. This finding is in line with Matisoff & Johnson (2017) who find that long term compensation incentives had minimal impact on increasing residential solar adoption. Net metering represents a compensation strategy that impacts the solar owner in

the long run. It would seem reasonable to suggest that solar adopters consider their long-term compensation strategy, but as is nature with most future compensation, the present may matter more. Factors such as the current retail rate of electricity and personal political sentiment have a more immediate and direct impact on the process. Additionally, outside of having direct insight into others' experience with net metering, a potential solar adopter may not be able to gather a complete picture of the potential positives and negatives of such a policy.

Lastly, I find a lack of evidence to support the notion of an income cross-subsidy. Higher levels of residential solar are associated with larger economies as measured by real GDP, but they are also associated with lower values of owner-occupied housing units. Assuming wealthier families own more expensive housing, this is evidence that the wealthiest populations are not those that are pursuing residential solar. More importantly, I found that mean income of a county had a zero impact on average in the overall model and a negative impact in the Probit selection model. If solar energy was truly pursued only by the wealthiest households, then we would expect to see more installed solar as the mean income of a county and the value of owner-occupied housing units increases. These factors considered, suggest that residential solar is likely pursued in counties that are characterized by the middle class. If this is true, the issue of cross-subsidization, a situation where wealthy individuals are subsidized by low-income individuals, is muted.

A limitation of these findings is the reliance on aggregate county level data where kurtosis, or distributional tail effects, are possible. It may be that households that install solar in these counties are the wealthiest households, but the data are driven by those who do not have solar. Future research with more granular data can further examine this issue.

REFERENCES

Ambrose, M. (2019, October 02). *FY20 Appropriation Bill: National Science Foundation*. American Institute of Physics. https://www.aip.org/fyi/2019/fy20-appropriations-bills-national-science-foundation

Bureau of Economic Analysis. (2020a). CAGDP9 Real GDP by county and metropolitan area [Data table]. Available from Bureau of Economic Analysis Website: https://apps.bea.gov/itable/index.cfm

Bureau of Economic Analysis. (2020b). MARPP Regional Price Parities by MSA [Data table]. Available from Bureau of Economic Analysis Website: https://apps.bea.gov/itable/index.cfm

Bipartisan Budget Act of 2018, Pub. L. 115-123; 132 Stat. 64 (2018)

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators. *Journal of Applied Econometrics*, 1, 29-53

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata.* College Station, Texas: StataCorp, LP.

De Groote, O., Pepermans, G., and Verboven, F. (2016). Heterogeneity in the adoption of photovoltaic systems in Flanders. *Energy Economics,* 59, 45-57. Available at: http://dx.doi.org/10.1016/j.eneco.2016.07.008

Energy Information Administration. (2020). *Electric Power Monthly with Data for November 2019*. Retrieved from https://www.eia.gov/electricity/monthly/archive/january2020.pdf

Government Accountability Office. (2018). *Climate change: analysis of reported federal funding* (GAO Publication No. 18-223). Washington, D.C.: U.S. Government Printing Office.

Greene, W.H. (2003). Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis*, 19, 179–190. https://doi.org/10.1023/A:1022853416499

Kavlak, G., McNerney, J., & Trancik, J., (2017). Evaluating the causes of cost reduction in photovoltaic modules. *Energy Policy*, 123, 700-710. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2891516

Lew, D., Brinkman, G., Ibanez, E., Florita, A., Heaney, M., Hodge, B. M., Hummon, M., Stark, G., King, J., Lefton, S. A., Kumar, N., Agan, D., Jordan, G., & Venkataraman, S. (2013). *The western wind and solar integration study phase 2* (NREL/TP-5500-55588). Golden, CO: National Renewable Energy Laboratory.

Matisoff, D., & Johnson, E., (2017). The comparative effectiveness of residential solar incentives. *Energy Policy*, 108, 44-54. Available at: https://doi.org/10.1016/j.enpol.2017.05.032

MIT Election Data and Science Lab, (2018), County presidential election returns 2000-2016, *Harvard Dataverse*, 6, Available at: https://doi.org/10.7910/DVN/VOQCHQ

National Bureau of Economic Analysis. (2017). *CMS's SSA to FIPS CBSA and MSA county crosswalk 2017* [Data file]. Available from National Bureau of Economic Analysis website: https://data.nber.org/data/cbsa-msa-fips-ssa-county-crosswalk.html

National Research Council. (2010). *Advancing the science of climate change*. Washington, DC: The National Academies Press.

National Renewable Energy Laboratory. (2013). *Lower 48 and Hawaii PV 10-km Resolution 1998–2009*. [Data file and code book]. Retrieved from https://www.nrel.gov/gis/data-solar.html

National Renewable Energy Laboratory. (2018). *U.S. Solar Photovoltaic System Cost Benchmark: Q1 2018*. (NREL/TP-6A20-72399). Golden, CO: National Renewable Energy Laboratory. Retrieved from https://www.nrel.gov/docs/fy19osti/72399.pdf

StataCorp. 2019a. *Stata 16 Base Reference Manual*. College Station, TX: Stata Press.

StataCorp. 2019b. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

United States Census Bureau. (2018). *National Counties Gazetteer File.* [Data file]. Retrieved from https://www2.census.gov/geo/docs/maps-data/data/gazetteer/2018_Gazetteer/2018_Gaz_counties_national.zip

Wilson, P., (2015), The misuse of the Vuong test for non-nested models to test for zero-inflation, *Economics Letters*, 127, 51-53. doi: 10.1016/j.econlet.2014.12.029

Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data* (1st ed.). Cambridge, Massachusetts: The MIT Press.

Yu, J., Wang, Z., Majumdar, A., & Rajagopal, M. (2018). DeepSolar: a machine learning framework to efficiently construct a solar deployment database in the United States. *Joule, 2(12),* 2605-2617. https://doi.org/10.1016/j.joule.2018.11.021